Corpus-based Research in Computational Comparative Literature

Christine Ivanovic
Andrew U. Frank

1 The Task of Comparative Literature

Comparative literature designates an academic discipline within the humanities characterized by its specific way of investigating literature (as well as other media) by comparison. Comparison however is widely understood as one of the most basic cognitive practices, and nowadays it can be performed even by machines; yet it has proven indispensable for all scientific approaches in all fields of research. Comparative literature as a discipline has to clarify why it considers comparison its basic task by definition, as well as explain the means and the goals of its specific approach.

As an academic discipline, comparative literature emerged relatively late. It arose as an attempt at comparing the observations and results of at that point already highly sophisticated philogical works in different fields of national literatures [References]. Comparative literature thus implies not only a comparison of texts, but a reconsidering of long established categories that are held to be determinative for literary texts, such as language, genre, nation, and others.

By investigating texts in different languages, or from different cultural, national, ethnical, religious, historical, or even media contexts, comparative literature is conceived of as a way of either transcultural or crosscultural examination. It asks for qualities of texts which are considered comparable throughout all literatures, without regards to (cultural) space and time, as well as for the manifold ways of transferring (cultural) concepts through space and time. Comparative literature is therefore as much about texts and concepts as it is about contexts and relations (which also includes investigation of the relationship between literature and other media, e.g. arts or film).

To compare in comparative literature means to simultaneously consider texts which must belong to different systems, and which are related to or referring to different contexts. Additionally, the comparison is also determined by the whole spectrum of theoretical approaches typical for the humanities, for example aesthetic and/or literary theory, history, sociology, political, cultural, religious or gender studies, and many others.

The analytical goal of comparative literature is not the hermeneutic interpretation of one single text, but (a) the determination of constant characteristics of

(literary) textuality, (b) the determination of historical changes of (literary) textuality, (c) the discerning of manifest proof of cultural contact (dependencies), and (d) the evaluation of literary representations and their references to "the (social, economic, political or religious ..) world". In other words: Comparative literature investigates texts simultaneously in multiple languages in order to better understand cultural progress and cultural systems as they are represented in texts.

Comparing texts is the exclusive means considered appropriate to reach this goal. To that end, several steps must be performed one after the other: (a) determination of text characteristics; (b) investigation of the same categories of text characteristics in multiple texts in different languages; (c) measuring of the specific qualities of text characteristics of the investigated texts; (d) evaluation of similarities and differences of characteristics of texts formulated in different languages and situated in different contexts; (e) evaluation of the different reference frames of the text or the text characteristics, respectively.

2 Corpus-based research in Comparative Literature

Comparative literature is based on comparing selected characteristics in a specified number of texts. As a starting point, each study in Comparative literature has to determine (a) the selection of texts to be compared, and (b) the characteristics to observe. Comparative literature is very much corpus-based, i.e. based on a quantitatively discernible number of texts chosen by single sets of qualities. The analysis depends on the type and number of the selected and analyzed texts as well as on the type of the characteristics observed.

If a study aims at investigating e.g. the importance of Dante Alighieri's *Divine Commedy* in respect to James Joyce's *Ulysses*, the comparison can be restricted to those two texts. But it might also be argued that including Homer's *Odyssee* as well as Vergil's *Aeneid* among other reference texts important to both of them into the corpus would also be advisable.

If another study investigates e.g. the typical subject of the "European Novel", the number of texts to be included as well as the selection of characteristics are less easy to determine [9]). The decision depends on prior concepts outlining which texts are considered to be a "novel" and which texts belong to "European literature". Likewise, the number of texts which by definition should be included may exceed the texts which are de facto accessible for investigation. As in the case of the study on Joyce's novel, it must be explicitly argued which texts are included and by which qualities they are chosen over other possible texts. The same goes for studies in typology, image studies, influence studies and other fields of comparative literature.

Comparative literature has to be corpus-based when it aims at generating (universally) valid results from the comparison of texts. But in order to conduct valuable corpus-based research, it is not just important to construct well-argued text corpora suited for this research; investigating the categories and concepts which

constitute a corpus in literary studies is also task to be solved by the discipline.

3 Computational corpus-based research overcoming the current limitations for comparative literature by scale

The tasks of comparative literature exceed the capacities of one single investigator. In theory, all (literary) texts ever written are potentially the object of comparative literature. In practice, however, only a limited selection of texts is evaluated. Traditional comparative literature is limited by the abilities, time, and resource constraints of the researchers. Proficiency in many different languages, cultural systems, and literary traditions is an inevitable precondition for a scholar engaged in comparative literature. But even if the scholar is broadly educated or working in a well-prepared team of investigators, and even if the study is restricted to a relatively small number of languages, as it was the case in our example of the "European Novel", the number of literary texts which should be included may be abundant, and beyond what is manageable even in a lifetime of a researcher. The effect of such restrictions is the approximative nature of all such investigations, based on a more or less personal and biased choice and a relatively small number of texts evaluated; that is the regular case, not an exception. As comparatists, we most often conduct (case) studies on a very limited textual base - with accordingly limited results.

For those reasons research in comparative literature only recently turned to considering a corpus-based computational approach in comparative literature. It makes explicit, in the construction of the corpus and in the identification of observations, which texts and which characteristics will be used for comparison. One difference to current practices is the ways in which the selection of texts and characteristics are being documented. The extension of textual comparability by enlarging the number of characteristics to be compared might be another advantage. Computational methods are intended to overcome current limitations in comparative literature especially in two regards:

Problem of subjective selection and circular logic The subjective selection by the individual scholar is hard to justify objectively. The selection criteria and the result of a study are often directly related. For example, in a study to describe the characteristics of the "European Novel", decisions on text selection are necessary. The choice of which texts to include anticipates the results of the study. This is especially problematic when following the classical credo of comparative literature: researchers are expected to only analyze texts that they can read in the "original language".

In order to perform a study on e.g. railway novels, one would recollect all the texts encountered so far mentioning railways, and then search for more texts which might fit. Then a selection of the most appropriate pieces is subjected to closer examination. This mode of operation is corpus-based, but the corpus is not clearly

determined. It consists of the texts already known by the investigator, and of texts searched for more or less systematically. Using this method, it is very difficult to develop or argue the criteria which lead to the choice of the texts used for closer examination.

A corpus-based computational approach would evaluate texts not through the personal choices of the investigator, but determined by the construction of the corpus. If using e.g. the Austrian Academy Corpus in the investigation of railway novels, a selection of over 6 million tokens from representative German language texts between 1848 and 1989 would be analyzed. This enables valid statements concerning this corpus. The findings can be compared with results based on another comparable text corpus in one or more other languages, and reproduced by other investigators. Researchers are not required to read all texts in all languages, but nevertheless will be able to compare them. It is evident, that with another corpus different results could be obtained. How to conduct a valid, i.e. a unbiased corpus is worth further studies. For now, if results for similar studies executed with different corpora coincide one might assume that the corpus is unbiased (or both corpora are similarly biased).

Problem of capacity Any scholar can only analyze a limited number of texts, which limits comparative studies to very small subsets of the vast number of literary texts ever written. This problem is especially pertinent to the aim of comparing texts in virtually all existing languages. Traditionally comparative literature rarely includes texts in so called "small" languages (e.g. Finnish, Urdu). Comparisons between texts in non-related language systems (e.g. Chinese to Arabic) are also conducted far less than those between "European" languages (e.g. English to French).

Intended Solution Currently, the vast majority of digital literary studies is focused on digital editing. There is a rising number of national initiatives to digitize huge text corpora in the respective languages. There are also Google Books as well as crowdsourcing initiatives like Project Gutenberg whose aim is to digitize more and more texts in many languages. As comparatists we should be prepared to harvest these treasures in order to seriously evaluate "global" or "world literature". Computational methods help to overcome some of these limitations:

- A digitally processed corpus includes a fixed, known set of texts.
- Digital text processing makes it possible to collect and analyze large text corpora / large amounts of data in any language, far beyond what an individual researcher can digest.
- Conclusions in a study can be corroborated in repeated studies and further checked with larger corpora.
- The selection of the corpus is logically separated from the conclusions.

- Results of corpus-based research are valid "for all texts included" and justify negative statements ("no text in the corpus has property x") - comparable negative statements are not justified without a fixed corpus.

4 Challenges for corpus-based computational comparative literature

The challenges for a computational approach to comparative literature research are twofold: (i) design an efficient structure for the corpus and its maintenance; and (ii) construct the computational methods to evaluate the texts in the corpus systematically and report results suitable for analysis.

The overall design must make it easy and fast to add a text to the corpus, or add or change one of the methods used for the evaluation of a text. The collection of results must be maintained up to date, which means: (a) if texts are added, all the methods must be applied automatically to them; and (b) if methods are added or changed, the new method must be applied automatically to all texts. Such a system of maintaining a literary corpus can be realized with today's information technology.

Construction of the corpus Adding a text to the corpus means to bring it into a form the automatic methods for evaluation can use it. Initially, for each text in the corpus must be available:

- bibliographic detail, at least a BibTex entry from which author, date of publication, edition etc. can be retrieved but better a description following the Dublin Core standard,
- the text with markup for the text structure, non-literary text parts etc.
- a treebank tagged version, preferably with dependency trees (parse trees), coreferences and named places recognized.

The markup of the text identifies the structure of the text in parts, chapters, paragraphs (or whatever a text uses), and the layout of the text on pages. It is also necessary to identify the title page, author names, and other material which appears in a printed book and identifies the source of the text but is not part of the literary text properly and should not be included in the analysis. The text should be in UTF-8 encoding to allow for non-alphabetic language texts, including Chinese, Arabic etc., and processed with a Natural Language Processor to get parts of speech tags, named entities etc. recognized. Markers are to be placed into the text such that the tagged text can be connected with the original text structure. Eventually, sentiment analysis should be included.

The marked-up text and the output from the speech tagger are integrated into a single framework. We currently experiment with RDF[8], which seems to fulfil

the requirements and is able to deal with the amount of data expected. Our observations indicate that for each word in a text with rich linguistic annotations, we obtain 10 RDF triples; for a literary text of 100,000 words we therefore obtain a million triples, or for a corpus of 10,000 books, 10 billion triples, which is well within the possibilities of today's RDF storage systems (benchmark results report that 1 trillion triples load in few hours [1]). Response times for different types of queries are acceptable even for very large data collections (for example dbpedia with 3 billion triples[2]), and hardware requirements are within the reach of a properly funded project (the test configuration for the above mentionmed project was acquired for Euro 70,000 in 2012).

Collection of methods The pertinent literature reports a large number of methods to evaluate texts, many of which are available on the web. Corpus-based examination of literary texts allows us, for example, to identify citations or reuses [7]. It serves for authorship attribution, for recognition of author gender/gender difference [10], or for the identification of literary movements [1]. It supports pattern discovery and similarity computation [11]. It has also proven useful in detecting emotions[5], or in knowledge mining[4]. If a well-structured corpus is available, new methods can be realized, for example:

- compare temporal sequences of actions with the sequence of narration in the text,
- spatial reference in the text: mapping the named places, amount of "movement" in the text, use of locations to convey e.g. sentiment, use of locations to link persons, etc.
- number of persons acting in the text and distribution of actions to them (hero-centered text, Bildungsroman, ..),
- vocabulary: literary sense vs. metaphorical, size of the vocabulary.

Each method produces for each text a value - True/False, a count, or a statistical value. For each text a vector of values represents the result of applying all methods to this text; this vector characterizes the text in the corpus.

Cluster analysis Applying all methods to all texts produces a matrix of evaluation values; a vector of values characterizes each text. Cluster analysis applied to these vectors indicates which texts are in some respect similar to each other. The resulting similarity structure can be mapped and analyzed in two directions, for example:

[1] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/

[2] https://en.wikipedia.org/wiki/DBpedia

- Reconstruction of traditional genre attributions for the texts. We ask which evaluation methods give similar results for a group of texts which are considered of the same genre, and gain an understanding of the connection between traditional literary research approaches and the computational methods
- Reconsidering literary genres and groups: Analyzing groups of texts which appear similar on some evaluation methods but are not typically thought of comparable (for example, compare the *Arabian Nights* with Boccaccio's *Decamerone*).

From the analysis of the results, new and refined ideas for evaluation methods will result, refining the computational characterizations of texts.

5 Examples for corpus-based computational comparative literature

Moretti's *Atlas of the European Novel* In comparative literature, computational corpus-based text examination is still in its infancy. Franco Moretti has been one of the pioneers of the field. His "Atlas of the European Novel" [9] from 1999 was one of the first studies that outlined some of the possible directions comparative digital literary analysis might take. It was the starting point of corpus-based research methods still to be developed in computational comparative literature. Moretti has developed parameters by which the analysis of a large number of texts could reveal connections between variables that have so far never been related to one another, for example: language and distribution of a novel in space and time (including distribution as a translation), diegetic space of a novel, characteristics specific to the sub-gerne (adventure novel, ghost story, Gothic novel...).

Novel networks In the scope of a current project conducted on the Austrian Academy Corpus (AAC hosted at the ICLTT/Austrian Academy of Science: **(author?)**, **(author?)**) as well as using other corpora, we plan to analyze the connection between the development of the railroad network from the second third of the 19th century until today in relation to the structural development of the (European) novel.

Multilingual corpora With new computational methods, comparing texts in multilingual corpora becomes possible. By using the approach outlined above of "all methods for all texts", structural text characteristics in particular can be compared beyond language borders. Possible research questions include:

- the evaluation of language characteristics, e.g. a comparative analysis of the proportion of multilinguality in texts,
- stylometry irrespective of language, for example, which characteristics of a style are preserved when translations are published [6]

- the structure of character depictions: the comparative analysis of the number of characters in texts in relation to the texts time of writing, text category, and text complexity; comparative analysis of character depictions (how they are being described through indirect speech, through attributes, through concrete action, through movements in space, etc.),
- the grounding of the plot through identifiable place names or historical dates.

6 Conclusion

Corpus-based research in Computational Comparative Literature has a supportive function for established approaches in Comparative literature. It allows for systematic evaluation of a large number of texts that goes beyond established questions. The aspects of the proposed technical solution make it even suited for the comparison of multi-lingual text corpora, as the same methods can be applied to all texts irrespective of language. In the terminology of software engineering, a corpus is an abstract data type (object) with methods to

- add and remove texts,
- inquire about the texts included, and
- query the corpus with a flexible query language.

For each discipline the units of texts which are annotaded and the annotation content must be adapted to the specific tasks; for many disciplines, an automated production of annotations is desirable to achieve reproducible results. The annotations inserted into the text determine the queries which can be asked - or reversed, what must be annotated follows from the expected queries.

The core is the construction of a large corpus of literary texts with sub-corpora and a related set of computational analysis methods. Each text is systematically analyzed with all methods and we obtain for each text a comparable collection (a vector) of characterizing values. Such a combination of corpus and method collection is feasible with today's information technology, and the text resources are available on the web. All methods for all texts!

Literatur

[1] Diego R. Amancio, Osvaldo N. Oliveira Jr., and Luciano da F. Costa. Identification of literary movements using complex networks to represent texts. *CoRR*, abs/1302.4099, 2013.

[2] Hanno Biber and Evelyn Breiteneder. Fivehundredmillionandone tokens. loading the AAC container with text resources for text studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 1067–1070, 2012.

[3] Hanno Biber, Evelyn Breiteneder, and Karlheinz Mörth. Words in contexts: Digital editions of literary journals in the äac - austrian academy corpus". In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[4] Amedeo Cappelli, Maria Novella Catarsi, Patrizia Michelassi, Lorenzo Moretti, Miriam Baglioni, Franco Turini, and M. Tavoni. Knowledge mining and discovery for searching in literary texts. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.

[5] Daniel Dichiu, Ana Lucia Pais, Sunita Andreea Moga, and Catalin Buiu. A cognitive system for detecting emotions in literary texts and transposing them into drawings. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10-13 October 2010*, pages 1958–1965, 2010.

[6] Umberto Eco. *Dire quasi la stessa cosa*. Bombiani, 2003.

[7] Jean-Gabriel Ganascia, Peirre Glaudes, and Andrea Del Lungo. Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, 29(3):412–421, 2014.

[8] Frank Manola, Eric Miller, Brian McBride, et al. RDF primer. *W3C recommendation*, 10(1-107):6, 2004.

[9] Franco Moretti. *Atlas of the European novel, 1800-1900*. Verso, 1999.

[10] Urszula Stanczyk. Recognition of author gender for literary texts. In *Man-Machine Interactions 2, Proceedings of the 2nd International Conference on Man-Machine Interactions, ICMMI 2011, The Beskids, Poland, October 6-9, 2011*, pages 229–238, 2011.

[11] Masayuki Takeda, Tomoko Fukuda, and Ichiro Nanri. Mining from literary texts: Pattern discovery and similarity computation. In *Progress in Discovery Science, Final Report of the Japanese Discovery Science Project*, pages 518–531, 2002.